Biography of Data: A Societal Level Perspective On Data Quality

Biography of Data: A Societal Level Perspective On Data Quality

Research in Progress

Farjam Eshraghian[1,a], Ashley D. Lloyd[1], Stephen A. Harwood[1]

[1]Business School, University of Edinburgh

[a]Farjam.Eshraghian@ed.ac.uk

Abstract

The prevailing view to the data lifecycle is the same as the view to the process of manufacturing of products and it is called data production process. Different data roles which have been considered in this process are data collector or generator, data custodian and data consumer. Lee (2004) suggests that if these roles have adequate contextual knowledge, the quality of data improves significantly. At the societal level, "psychic distance" between different data roles needs to be accommodated. We argue that networked-view (Law and Singleton 2005) to the process of data production helps to have better understanding of how these roles can affect data quality at the societal level. According to this view an independent data role can be assumed for the technology which is being used in the process while its role is generally considered to be incorporated in other data roles. Leonardi (2010) argues that digital technology has materiality, hence, we decided to consider a role for information technology in this process which leads us to the technology vendor's indirect role in the process of data production. We argue that the data consumer should understand other data roles' perception from his/her context is crucial for improving the quality of data when one or more data roles are from different societal groups (countries). Drawing on Kopytoff's biography of things, we suggest biographical perspective gives complementary insight into data quality issues.

Keywords:

Data quality, Biography of data, Iran, Psychic distance, Societal Level, Context, Contextual knowledge

Introduction

> 'social distance, and leading on from there into bigger societal things, is very, it's really interesting, as part of the problem of doing a data migration is actually doing the job as opposed to looking at the data'
>
> John Morris, author of Practical Data Migration, 2006, interviewed 2014.

This comment about data quality reveals the importance of social distance interpreted as the 'psychic' distance between cultures, is a factor that needs to be accommodated when considering quality of data at the societal level. It has been generally assumed in the literature that countries' specifications have been granted into the notion of the context. We argue that this heterogeneous view of business context (assuming many unique contexts) regarding data quality may prevent understanding of similar problems which exist across different contexts. Although a wide range of contexts are imaginable, the extent of uniqueness of problems which are raised by a specific context are not completely open to many possibilities (Monteiro and Rolland 2012; Orlikowski 2000). We argue that problems with the quality of data at societal levels (country level) can be solved by similar methodological approach.

Today, researchers agree the importance of data quality and that data must fit or be tailored to the data consumers' context of use if it is to enhance practice and have impact (Lee et al. 2007; Lee 2004; Pipino et al. 2002; Strong et al. 1997; Watts et al. 2009). Capturing context is also an important aspect of empirical research on data quality that allows the pathways to impact to be mapped. Lave (1993) defines context as a relational construct that represents interactions and relations between people and the social world surrounding them. She also notes that these relations can be considered very solid ones from a historical perspective, cutting across

multiple instances/circumstances over time.

Lee (2004), reviewing studies adopting a contextual approach, argues that studies concentrating on more than one context have greater value. She suggests that these new studies should include both political and cultural contexts. Zhu et al. (2012) conducted a comprehensive literature review of quality of data and different frameworks and have also concluded that researchers considering emerging trends in the field of data analysis, such as big data and the harvesting of unprecedented volumes of data from social networks, should consider conducting research on data quality at different societal levels.

Whilst it is generally safe to argue that more data and more contextual data provide more pathways to impact, the aim of this paper is to focus here on the extent to which the quality of data, when measured in its use context, is associated with different societal levels of context.

This paper is structured as follows, the first part discusses about the concept of context, the manufacturing view to the concept of data and data production and the network-perspective of data production process considering different data roles. The next part outlines the methodology of this study. The third part provides brief description of the history of Computing technology and web in Iran. The case study which has been chosen for this paper is from the context of this country. In the discussion section, it is argued that to what extent it is important to consider the technology which is being used in data production process as a data role (actor). It is also emphasised that understanding the technology vendor's perception of data consumer's context for data consumer is critical to improve data quality when one or more data roles are from different societal groups (countries). The conclusion section underlines the biographical view to the concept of data and to the evaluation of data roles' perceptions of the data consumer's context.

Background and Theoretical Framework

During this study, we have accepted the data production process or information manufacturing framework that is widely used by the researchers in the field of data quality (Ballou et al. 1998; Lee 2004; Lee and Strong 2003; Strong et al. 1997; Wand and Wang 1996; Wang et al. 1996). According to this view, the stages of data production are data generation or collection, data storage in databases and data processing for consumption, and all the value added is measured in the context of a specific set of data consumers, allowing an analogy to a product manufacturing processes to be employed. As manufacturing processes convert input materials to a product that is required by the eventual consumer, a piece of data must be converted to what is required by data consumers. Most studies consider the process to map to 3 different roles: data collector or generator, data custodian and data consumer, occupied by people or resources that generate, administer and consume data. Lee and Strong (2003) argue that if the people in these roles have significant knowledge about what is data about, how data is collected or the procedure of data collection and why data is collected (knowledge about context and principles) it improves the quality of data emerging from the 'production' process.

Developing Lee and Strong's (2003) position, we argue that the data production process can be considered to delineate a network with different actors (data roles) including human and non-human actors (Law and Singleton 2005). This view helps with analysing the process of shaping a stable network with different actors that have various interests but align to produce a stable network. Should the network shaping be unstable, and hence unsuccessful, the network-perspective is also helpful in identifying the reason (Akrich 1992; Walsham and Sahay 1999). All actors' existence must be visible or sensed by other actors in the network, whether an actor is a human object or non-human (Law 1987). For example, this view explains why the quality of secondary data is held to be lower than the quality of primary data as in the case of secondary data the interests of data consumer as an actor need not be aligned with the interests of other actors such as data collector. Thus the network which has been shaped is a loose network, with different actors trying to inscribe their interests into the network while it is being shaped.

During this shaping process the role of technology, especially computing resources, is often neglected as an actor. Law and Singleton (2005) go further and remind us that an actor should not necessarily be considered a human or solid object, but might be considered something like a fluid, or volume, or set of configurations (such as technological configurations). Although most studies in the context of data quality have considered a role for technology or technological configuration, it has been largely subsumed into other human roles such as 'data custodian'. In most cases, the technology role has been considered to be implicitly aligned with other roles and the context in which data production process occurs, whilst for social scientists digital technology has a more distinctive materiality, as Leonardi (2010) notes:

> *their 'materiality' is determined, to a substantial degree, by when, how, and why they are used. These definitions imply that materiality is not a property of artifacts, but a product of the relationships between artifacts and the people who produce and consume them.*

Leornardi's description emphasises that both technology vendor and technology user impact materiality, thus, the technology vendor's perception of the technology user's context can be a determinant of the critical role performed by a technology in the context of use. This argument can also be extended to the data production process. Whilst Lee and Strong (2003) emphasise the contextual knowledge of the different data roles, the missing link here is the role of technology and the indirect role of the technology vendor's contextual knowledge.

Fleck (1994; 1999) introduced the notion of "innofusion" and noted that innovations, especially technological artefacts, do not diffuse directly after a development process to end-users as final products. He observed that development and feedback processes are simultaneous. The mutual relation between developers and users leads to inscription of users' interest into the final product. Williams et al. (2005) call this interaction "social learning". Development can therefore be considered shaping a network of different actors with different interests and power relationships (Latour 2005; Torchiano et al. 2011). Where interests conflict, power is a good indicator of how much their respective interests are going to be inscribed into the network, while the extent of conflict shows the looseness of the network that has been shaped (Torchiano et al. 2011). The failure of traditional 'technology push' approaches to development to conceptualise how products will be actually be adopted and shaped by users shows both the power of the vendor and that the different interests of a wide range of users have not been taken into consideration during the development process.

As the diversity of societal groups within the network increase, either geographically or culturally, there is no automatic shift in the power relationship between vendors and users, hence this divergence is likely to increase. This issue focuses on the question of how data roles originating from different societal groups impact the shape of the network and how well their different contextual knowledge is inscribed. Although globalisation has brought together people from different countries and societal groups together especially in the context of product markets which get more global every day, they have not left their culture, language and other specific characteristics that differentiate them from other societal groups behind to join the global village. Thus the answer to this question is important for practitioners and may help build significant links between the literature of data quality and other relevant domains. While differences are known to have serious consequence in the realm of international business, expressed as resistance to change and disruption to business practices (Shenkar 2001), it also points to the potential value of considering difference during data sharing between parties from two different societal groups. Levitin and Redman (1993) note that, in a data lifecycle, data creators are not necessarily the users and consumers of the same data they have generated. They also stress that a predefined view is necessary to collect, create, modify and work with data - in other words, this view determines what kind of data is acceptable, how it is gathered and how it is going to be consumed. Data users have their own views while using this data and their views are not necessarily the same as those of the data creators. The consequence of this discrepancy between views can lead to some data quality issues in practice (Levitin and Redman 1993; Redman 1998) and hence we argue that when data roles from different societal groups participate in data production process, their perceptions of each other's views and contexts have a critical impact on improving data quality.

Lee (2004) conceptualises context as the underlying domain and setting under study, which includes the related functionalities and occurrences. She notes that context differentiates and shows different relations between contents, what is inside, and settings. The importance of the role of context in data quality studies has been stressed by many scholars (Cao and Zhu 2013; Lee 2004; Strong et al. 1997; Wang et al. 1996), with Strong et al. (1997) underscoring the concept of context to encompass different aspects of data quality in "data consumers' task context". Lee (2004) and Lee et al. (2007) develop this further by decomposing context into the following categories to position it more centrally in the domain of data and information quality: goal, paradigm, role, place and time. Goal is the reason for which a piece of data is generated or collected, stored and used. Paradigm refers to the collective values, ideologies and principles which people who are engaged in the process of data creation, storage and consumption hold. These values and principles may be accepted among certain societal groups in a specific period of time and hence can affect how problem are seen and solved by people from these groups in different ways as the network evolves from one epoch to another. In the literature of data quality there are three different roles considered over a data lifecycle: data collector, data custodian and

data consumer. Time and place become aspects of context that refer to the specific location within which and specific period for which a piece of data is consumed.

The notion of "psychic distance" applied to the impact of ICT infrastructure development on markets (Lloyd 2005) can be useful more generally to perceive context at the societal level (Sousa and Bradley 2005; Sousa and Bradley 2006): *"it is the individual's perception of the differences between the home country and the foreign country that shapes psychic distance concept"*. We contend that considering psychic distance between different data roles is important. Lee (2004) states that practitioners solve problems with data quality by understanding "the work context" and try to make "a sense of situation as a whole". In this study we extend Lee's conceptual frame and argue that, in the societal level, when one or more data roles are from different societal groups (countries) it is crucial for the data consumer to understand how his own context (especially paradigms such as culture, language etc.) is perceived by other data roles.

Methodology

The societal context for this study is Iran, and probed via a single technology case study. The company's core technology, its website, is operated to sell products online. The company was founded by 3 friends in 2005 and now employs 30 people. Two of the founders were interviewed, one via Skype and the other via email. The company employs three Search Engine Optimiser experts, and two were interviewed via Skype and e.mail. Company selection followed a detailed review of the history of computing technology and Internet use in Iran to get a general framework to structure questions about societal characteristics of digital technologies.

English is Not The Language of Iran! Establishing A Biography Of Computing Technology in the Context Of Iran

Parhami (1997) in his article considers 5 different stages for the history of computer systems in Iran. The first stage (1950-1960), appearance, was punctuated by the entrance of first computer systems in Iran in 1950. These first systems were electro-magnetic devices that processed data and were imported to be used exclusively in the oil industry (Feyzi 1985). The Iranian Actuarial Department started using the newer models and the first Iranian modern census was the result of applying these new machines in the context of actuarial activities in Iran (Feyzi 1985). In 1957, International Business Machine Corporation (IBM) was the first American company active in computer hardware industry to open an office in Iran and provided related equipment such as punched cards and devices for reading and writing to those punched cards for the Iranian Actuarial Department during the time of first modern census (Dadeh Pardazi Iran (dpi)). At the beginning of the 1960s, Persian Computing entered into its second stage which we typically recognise as the "development stage" (Parhami 1997), although Parhami notes that this stage may also be described as the "Contagion Stage" with competition among different organisations, especially governmental ones, promoting the acquisition of more computer hardware without establishing any requirement for a feasibility study. As a consequence Feyzi (1985) notes that Iran took a 'trial and error' approach to Information Technology during this decade. Until the middle of the 1960s, the adoption of computer systems was very slow and was limited to especial sectors such as oil industry, banking section and some governmental department (Feyzi 1985).

The start of the 1970s coincided with increasing oil prices in the global market and the flow of oil money to Iran. Competition intensified among government departments and rich private companies were able to buy expensive and large-scale computer systems (hardware) and install them in the country (Feyzi 1985; Parhami 1997). This stage is referred to as the "Review Stage" in Persian Computing history (Parhami 1997).

"The huge increase in oil revenues" in the start of 1970s resulted in "subsequent [Iranian] government expenditure" (Dadkhah 1985) and procurement of IT facilities were no exception. During this time there was no regulation in place to establish standards and hence restrictions on companies to buy IT equipment from specific companies to specific standards and hence competition among different managers in a growing market with increasing budgets promoted the acquisition of more equipment for their organisations without necessarily having even the most basic knowledge about Information technology (Feyzi 1985).

Although, through this surge in oil prices, Iran's modernization accelerated in the middle of the 1970s, many parts of economy were running traditional systems to support agriculture and rural production and hence for many companies and organisations a computer system was a luxury. Relatively few companies, such as governmental organisations, military, and a couple of private organisations, had the money to pay for expensive computer hardware and equipment at that time (Sadoughi 1981).

These constraints were reflected in Iranian companies buying computer components, hardware, and software from foreign companies such as IBM and HP and recruiting more IT staff without paying necessary attention to supporting investments in infrastructure or to planning for future staff requirements given the availability of domestic professional human resources (Parhami 1997). It has been observed that many Iranian organisations were attracted at the very early stage by Information Technology as a fashionable investment, or as Wang (2010) defines it: "a transitory collective belief that an information technology is new, efficient, and at the forefront of practice". This behaviour was reinforced by the entry of multinational companies into the Iranian market, such as HP, Honeywell and Data General, who could see clear benefits of extending into Iranian markets, with IBM established as the market leader (Feyzi 1985).

The dominant application of these systems in Iranian organisations was batch processing rather than on-line processing, allowing treatment as a distinctive company operations that could be isolated from corporate processes rather than embedded and diffused. Few organisations were capable of linking this processing operation into Management Information Systems with only some banks and Iran Air (Iranian National Airlines) using on-line processing for its international ticket reservation, with backup systems in both New York and Frankfurt (Feyzi 1985).

Persian Computing in an 8-Bit ASCII World

Diffusion of computer systems accelerated following the unbundling of hardware and operating systems by IBM, leading to the lower-cost cloning of IBM PC hardware and the development of alternative operating systems such as Microsoft Windows. This move from a command-line operating system to a graphical user interface in the early 1980s made computers more accessible to non-technical people and cheaper to own. Export restrictions meant that this diffusion happened at a slower pace and, given income disparities, with a more limited scope in countries such as Iran.

A significant barrier to the domestication of this technology within Iran, and hence its diffusion among small and medium sized businesses, was the lack of support for the official language, Farsi, which had fundamental differences from languages such as English, French, and Spanish in terms of the required address space: whilst the first languages to be supported had character sets that exist within an 8-bit (or even 7-bit) address space, languages such as Arabic and Farsi required multiple bytes. Microsoft started adding the support of additional languages to the later versions of Windows but the support of Farsi wasn't added. This was highlighted as a critical factor in the decision by the user community to domesticate this technology by one interview respondent, an Iranian software engineer who specialised in software localisation:

> *Microsoft was an American company and because of sanctions against Iran they didn't want to invest in something which didn't have return for them at first.*

After the release of the first Windows version to support Arabic, two Iranian companies extended the character set to support Farsi. This temporarily solved the problem and became a de facto standard within Iran.

A decade later, Microsoft officially added support of Farsi to Windows but "*support in Microsoft products was not perfect [...] the Persian language has unfortunately merely been considered a variant of the Arabic language*" (Esfahbod 2004). For some years afterwards this problem "*affected the common practice and user experience of some finer details of Persian computing*" (Esfahbod 2004).

The history of computer system in Iran has been coupled with the developing support of the Farsi language by widely-used applications such as text editors, accounting applications and popular operating systems such as Microsoft DOS and Windows. A country's or societal group's language is an important differentiator of that group from others in how they use digital technology. The brief Persian computing history, here, confirms the central role of developing support for the Farsi language in the adoption of Information technology by Iranian organisations.

Web in Iran

The first Internet users in Iran were universities and public organisations in 1992. Then news agencies, households and private businesses started using the Internet widely by the beginning of the millennium, coinciding with the decrease of connectivity tariffs and appearance of Farsi websites. Hosting and administrating Farsi websites was at first a major challenge as most web technologies could not render the Farsi

text and had problem with non-Latin script languages. In addition to this, Farsi is written from right to left and showing a multilingual text (combining Farsi and English, e.g. scientific units) was an inefficient departure from established approaches to localisation. Although the World Wide Web became popular among households, especially younger generations, and some businesses before the support of Farsi, it was only used for necessary services such as instant messaging and emails. For using these services, Iranian used 'Finglish' - writing Farsi language using English scripts - to communicate via instant messaging and emails.

While it was considered a daunting task, it was possible to start Farsi websites and blogs by manipulating open source applications. The main problem was from the users' side when they opened these websites and saw meaningless glyphs instead of Farsi text. Although since the beginning of the 2000s some of browsers started supporting Farsi websites - for example Windows codepage 1256 could show most of Farsi text in to an acceptable degree - most normal and non-technical users did not have the skills to easily change their browsers' codepage to the requisite one.

This lack of language support combined with low investment levels in infrastructure were reported by respondents as contributing to relatively low levels of adoption of eBusiness in Iran compared to other countries. In fact the first eBusiness model to succeed in Iran was the 'free blog' providers who supported their business by showing adverts on the blogs, obviating any requirement for the direct commercial exchanges with consumers that fuelled the 1990s 'dot.com' boom in the 8-bit ASCII world.

Case Study: A Persian Online Shop ('A.Ir')

The business (which we will refer to as 'A.ir') currently sells goods and products online but was initially founded as a reviewing website for products in 2005. The range of products reviewed is wide but their principal focus was on products of interest to Iranian Internet users, such as different brands of Laptop, PC and mobile phone. Their reviews are considered to be useful and they ascribe this to their ability to attract a wide demographic of readers. The growth of the company was subject to competition and the core competitive competency reported by the founders is information management – a 'detailed and categorised' way of reviewing. After establishing a repeat readership, they started selling the products they had been reviewing. A lack of supporting infrastructure prevented Iranian banks from providing online payments for its customers, so the company accepted payment by cash with products delivered to a customer's door by Iran post. Although online payment became widely available through Iranian banks from 2007, the company still accepts cash because many Iranian users still don't trust online payment despite public promotion by the government.

One of the founders of A.ir noted:

> *One decade ago, there was no ecommerce in Iran except the services which were provided by a few online shops who used to sell books and CDs. Today, many people feel that this is very lucrative market to sell products online and the number of online shops is increasing every day. (Translated from Farsi)*

Market conditions have changed since the A.ir was founded. Although they have a proportion of loyal customers, they now have many competitors in the market and the market leader is an Iranian online store that was able to attract considerable amount of investment capital. As a consequence it was able to increase the range of products it sold, decrease the time of delivery considerably and as its market share grew, expended significant funds on advertising:

> *Customers used to be more loyal than they are now. Couple of years ago most of our customers bought from our website more than once but these days the competition is more fierce and we cannot afford to spend as much as some of our main competitors on advertisement. I think, it is much harder now, if you check, it may not be exaggeration to say that a couple of new online shops are being registered and entering into competition. Some of them could gain significant capital from somewhere to enter the market*

> *but most of them don't last long and we are not worried about them. But the few who stay in the market make it more competitive and if anyone wants to stay, it must find a way to stand out. Iranian users only keep one or two names of online shops in their mind not all of them. (Translated from Farsi)*

As already noted A.ir's strategy has been to promote its website by publishing review articles and engaging with users. Although this is ascribed to Air's past competitive success, the founder believes this strategy needs to be revised:

> *Iranian are good bloggers. I don't know why, maybe because the first localised web service was a blog provider but I should emphasise that they are really good bloggers. When we start as a laptop review blog, we could stand out. Nowadays, you can find tons of blogs and websites which review the products we review. Our reviews are good but if I want to be honest, some of their reviews are as good as ours. (Translated from Farsi)*

The first revision of A.ir's strategy was to recruit more technical bloggers who could review their products. A.ir wanted to improve the quality of reviews and to publish the reviews more often. They found out that Google was the key to improving their profile: if they could publish more reviews over the same period of time than their competitors they would improve their Google ranking and hence their market position.

> *Since the time we have started our website, we have known that Google is the key. But we could get the result we wanted by publishing detailed review about products, there was no such reviews in Farsi for Iranian users at first and we could easily rank very well by Google. We did not care that much about other search engine, I wonder who cares, every one, here in Iran, uses Google at least everyone I know uses Google. For example, if we were in China, we should have paid attention to Baidu or I don't know, maybe somewhere Yahoo is more important. Even, in other search engines we could rank very well but our analytics showed us that more than 80% of traffic came directly from Google. (Translated from Farsi)*

However, the success of this approach has been mixed according to the founders:

> *Well, it's hard to say it did not work but I should say, we did not get the result we had expected. (Translated from Farsi)*

> *Because of sanction on Iran, online retailers such Amazon cannot provide service for Iranian consumers and this fact had provided a niche market for localised businesses which had seen this opportunity earlier than others …… today, it is not as lucrative as before. When …. [the other founder] and I were searching for the sol5.ution to improve the rank of our website for Farsi keywords such as "خرید لپ تاپ دل" [buy Dell Laptop] or "مقایسه لپ تاپ" [Compare laptops] we could not really find very good reliable sources instead we turned to English online resources which were widely available online (Translated from Farsi)*

The founder's view was that where they do not have the required expertise in SEO, they should hire an expert to do it for them. The problem was that there were few, if any, experienced professionals in the Iranian job market

who had already worked as SEO professionals. The demand for this expertise emerged in Iran job market around 2008 and some of web developers had started studying about SEO based on English-language resources. The SEO articles available online at that time were simply translations rather than based on real experience in Farsi web. A.ir recruited one of the experienced web developers who was familiar with the principles of SEO, establishing a team of SEO experts that has now grown to 3 in number. We interviewed this initial SEO expert recruit, who still works for A.ir, and refer to him as SEO expert A:

> *Since 2008 when I started working here, I tried to apply what I had known about SEO to improve the ranking of the site in Google for some certain Farsi keywords about the products we review and sell and some general keywords such as "خرید" [shop, buy]. I had felt it at first that something was wrong what I had applied did not improve the ranking significantly after 2-3 months …. Everyone blamed me. (Translated from Farsi)*

SEO expert A noted that he felt he needed to review his resources in case that he had missed some logical step that had caused their effort to improve their rank to fail. Following review and revision their rank did not significantly change in their second attempt either:

> *Today after 6 years of experience, I can tell you that 2 important factors are really important while you try to improve a Farsi website ranking in search engines especially Google: Iranian users' habit and difficulties of searching a Farsi text. I think that search engines have not been designed for Farsi, at least…. I don't know about others languages but I think the tricks, which can be learnt from resources, work much better for English websites rather than Farsi websites. (Translated from Farsi)*

SEO expert A makes the interesting claim that the Google algorithm and Iranian users' habits change independently over time and are very loosely coupled, a combination that makes his team's job very hard. Top ranking for special keywords remains their main strategy to attract users to read their reviews, which lead some of the team to shop at their own website. SEO expert A explains some rather idiosyncratic aspects of Iranian users' search habits that are strongly tied to domestication of information and communication technologies:

> *For example, 6-7 years ago most Iranian users used to search in Finglish [using English alphabet to write in Farsi] for example "Kharide Laptop Dell" [Buy Dell Laptop]. But it has been changed since then. 2-3 years ago, most of them used to search using the combination of Farsi and English such as "خرید Laptop Dell" or "خرید لپ تاپ Dell" [Buy Dell Laptop], today most of them use in pure Farsi such as "خرید لپ تاپ دل" [Buy Dell Laptop] (Translated from Farsi)*

For this particular SEO expert, as Iranian users' searching habits change, this impacts the relative importance of specific keywords for A.ir over time and hence they must modify their 'SEO tricks' to get Google to rank them higher as the new keywords emerge. SEO expert A describes how nuanced this process is:

> *My guess is that this change of habit in this special case is due to change in people's texting habit. If you remember, couple of years ago mobile phones did not support Farsi very well, thus, people sent text in Finglish. I think their online behaviour was inherited from their texting behaviour. 3-4 years ago, most mobile phone supported Farsi and the government wanted to change users' texting habit. It has decreased a text tariff which is completely in Farsi considerably compared to a text tariff which contains*

> *English characters as well. I think this policy has changed the people texting habit and my guess is that it also affects how people search online. (Translated from Farsi)*

These observations highlight the difficulty of forecasting behaviour from past data, SEO expert A noted that only recording users' search pattern does not help A.ir that much because these are past data and they need to focus on what changes are likely to happen in future. A better understanding of the context of changes is likely to improve A.ir's ability to use the data to predict future behaviour, however at present A.ir takes a more reactive approach, placing priority on quick identification of emerging keywords to help assess potential changes in customers' searching behaviour:

> *Although we have some loyal customers who usually come and read our reviews, the majority of customers are searching for bargain and quality in exchange for the money they pay. They can easily buy from other websites. It takes usually 1-2 weeks for Google to index changes in our website. I don't want to say that users' searching behaviour is changing every day but we should react sooner to rank higher. If our website does not rank very well for a good keyword, our customers drop significantly. (Translated from Farsi)*

This aspect of A.ir's approach is no different to SEO within an 8-bit ASCII environment, but implicitly assumes that the same targeting process will have the same impact on ranking. SEO expert A however had prefaced this priority for A.ir with the observation that "search engines have not been designed for Farsi" suggesting that the pathways to impact are different:

> *The best example that I have is difference between "لپتاپ" [Laptop] and "لپ تاپ" [Lap top]. For a user, there is no difference between these two terms both are perceived as laptop. But for Google, these two terms are different. If one of them is more repeated in the articles and Meta tags, that page is ranked higher for it rather than the others. It is not professional to use both forms in an article from consistency point of view. If one of them is included in Meta tags while it has not been used in the article, Google understands it's a SEO trick and we should expect penalty which is to decrease our rank or in some cases it may eliminate your website from its index. (Translated from Farsi)*

He stresses that his team try to have clear insight into the customers and thinks that a successful SEO strategy is a proactive strategy rather than a reactive one. He notes that his team divide their website visitors into different groups and try to categorise their respective data differently:

> *It is more likely that users who are over 35 use "لپتاپ" [Laptop] while users under 35 use "لپ تاپ" [Lap top] to search on Google website. My guess is that people used to avoid blank space while spelling one word. Since 20 years ago, Iranian education system has asked students to put blank space in spelling a word where they have stop in the middle of pronouncing that word. Thus, depending of the education system to which visitors belong they use different form to search. (Translated from Farsi)*

-though the age of visitors cannot be tied unambiguously to a specific browsing pattern:

> *Google Analytics service provides rough information about our website visitors' age. (Translated from Farsi)*

- making it hard to operationalize concurrent sensitivity to both groups without incurring a 'double counting' penalty from Google as their algorithms are evolving based on different cultural norms, with regular updates to its search algorithm that A.ir must respond to very quickly:

> *It has happened that Google releases update and we should always expect it. No one from outside really knows Google ranking algorithm but some people try to guess it by trial and error or reverse engineering. It takes time. When it releases an update what we already knew about its ranking algorithm slightly changes. Some people publish their experiences with these changes in ranking algorithm online but it is not that useful for us. We should discover its consequences for Farsi web and make the data that ranks our website better according to this new algorithm online as soon as possible. (Translated from Farsi)*

Discussion

Past studies about data quality have split the data creation, acquisition and modification cycle from the use and consumption cycle (Lee and Strong 2003; Levitin and Redman 1993). The roles discussed earlier that were outlined by Lee and Strong (2003): collect, store, use/consume are presented as distinctive yet can overlap or indeed be the same person as organisations increasingly expose their back-office processes to the consumer. In the context of our study these roles are clearly distinctive from each other -in A.ir there are different layers of data collection and consumption. At the first layer, the data collector is Google Analytics, an algorithm which collects data about customers whose behaviour, especially their searching habits, are valuable data on which A.ir defines its future strategy, making the A.ir SEO team data consumers. At the next level, A.ir's SEO team provide metadata such as tags and the structure of the articles published on the website, which Google indexes according to an evolving algorithm to use the word repetition, tags, etc. to rank the website based. Lee and Strong (2003) talk about the notion of 'knowing-why' in the data lifecycle and note that if different data roles have enough information about the context of data which is generated or collected, the quality of the data increases. This has conceptual appeal in the case of A.ir and appears to apply very well.

The examples provided by SEO expert A clearly show that the A.ir SEO team have good knowledge about the context of the data they are either generating or analysing. They had clear ideas about what had caused a group of their visitors to use "لپتاپ" [Laptop] instead of "لپ تاپ" [Laptop], suggesting that they understood "what, sociologically, are the biographical possibilities inherent in" data (Kopytoff 1986). This helped them to predict how best to sell a new product by directing them to the form of spelling used by the most valued group of consumers searching for the product on Google. This knowledge of the 'possibilities' enhances the quality of customers' data for A.ir and provides more pathways for the knowledge to have impact in comparison to the same dataset used by their competitors.

A.ir's SEO team also appreciate that these two words are interpreted differently by Google. This is what differentiates them in improving data quality problems compared to their competitors. For A.ir, Google is one of the actors participating in the data production process and shaping a network with other data roles. Understanding how Google perceives the Farsi language as one contextual element of the environment in which A.ir works is competitively critical for A.ir. Thus, we might conclude that considering only the work context and using work contextual knowledge is not enough: A.ir could have been impacted negatively not due to its type of business but because a key societal context had not been perceived by one of data roles.

Google represents both an algorithm and a core infrastructure. Its materiality (Leonardi 2010) is determined by the people who code the algorithm, infrastructure and the data which is used to modify the algorithm dynamically (Orlikowski 2007). The psychic distance between people who are behind Google algorithm and Farsi language should be taken into consideration and does not appear be being addressed from the perspective of A.ir, even though (a small number of) Iranian engineers currently work at Google. From the A.ir SEO team's perspective, Google's materiality has not been shaped around the context of Iran and it has not been designed

to understand and interpret the complexity of Farsi language. A.ir have monitored changes and updates in Google algorithm according to the context of their data and Farsi language, establishing a biographical record that describes this contextual factor. Different updates of Google can have different implications for types of language that are not within the centre of gravity of Google's global markets. Thus, it is clear that capturing the differing perceptions of the evolution of the data consumer's societal context by different data roles, can help the data consumer to improve data quality.

In this paper we have adopted biographical approach to relate detailed and diverse insights into the data lifecycle and its quality issues at the level of a societal group. We refer to this perspective as a 'biographical' view to data or 'biography of data', echoing earlier work by Kopytoff (1986) who describes the biographical approach as follows:

> *In doing the biography of a thing [piece of data or dataset], one would ask questions similar to those one asks about people: what, sociologically, are the biographical possibilities inherent in its "status" and in the period and culture, and how are these possibilities realized? Where does the thing [data or dataset] come from and who made it? […] what are the recognized […] periods in the thing's "life", [data lifecycle] and what are the cultural markers [societal characteristics] for them? […] what happens to it when it reaches the end of its usefulness?*

This approach helps characterise and couple the perception of different data roles from each actor's societal context where there may be considerable psychic distance between one or more data roles and the evolution of this perception.

The biographical approach has already been applied to different fields of study by different scholars such as Marcus (1995) who called it a "biographical narrative" in relation to his methodology for adopting multiple locations to study different subjects in the field of anthropology, Leonardi and Barely (2008) applied this biographical approach to study how technology materiality evolves throughout the time when it moves to the boundaries of different societal groups and social settings, Williams and Pollock (2012) who applied this approach to study the development of enterprise packages as technological artefacts.

We argue that by applying a biography of data perspective to the issue of data quality, A.ir was able to overcome some of the data quality issues it has encountered. A biographical view of the evolving Iranian search behaviour and Google's perception of Farsi language has helped them to understand the reasons behind this loosely coupled and hence challenging evolution and thereby make more sense of their data. Without this approach they could not have used their customers' data effectively to design a strategy that has sustained their competitiveness. Lee and Strong (2003) argue that when all data roles know the context very well the quality of data improves significantly. In A.ir's case they know the context well but this appears loosely coupled to Google's own algorithm development and might be explained by an imbalance between Google's knowledge of the Farsi market context in their data role versus their knowledge of other, 8-bit, markets to which they are better placed to respond. A.ir has tried to solve this issue by biographically analysing Google over time and understand the impacts of its relative lack of knowledge of their societal context. The A.ir SEO team have thus been able to manipulate data they provide to allow Google to rank them more appropriately. The alternative of trusting Google to do this for them could have had serious competitive consequences, principally losing their market share and position as one of the top internet retailers in Iran. A contemporary example of this is related by Lazer et al. (2014) who explained Google's over-estimate of influenza diffusion in terms of a failure to consider the context of this data and how this introduces biases that might be controlled for.

The main implication of this study for future research is to test the validity of the 'biography of data' approach in the context of Big Data and unstructured data. The internet is presented as bringing together unprecedented volumes of data at very high resolution about behaviour in social contexts across the world along with access to scalable distributed analytical resources, however making sense of this data, and sensible predictions from it, require mechanism such as the biography of data to cope with the fact that this data is produced by different people from different countries with different data structures and societal context elements such as language and culture.

Conclusion

In this paper, we have argued that a new methodological perspective is necessary for data quality issues arising from a societal level which complements existing approaches such as the context-reflective view to data quality by Lee (Lee 2004). Different data quality issues arise from trans-organisational and societal perspectives and when data travels from one organisation to another in a different country and from one context to another one, we suggest that a biographical approach drawing on Kopytoff's (1986) biography of things, which we entitle the 'biography of data', provides new and complementary insights into data quality issues. Using this methodological approach, we argue that contextual data quality should be seen from the societal perspectives of different data roles when one or more are from different societal group (countries). Societal characteristics can evolve through time and a biography of data is well positioned to capture these changes in a way that preserves the value of the data. A brief review on Iranian (Persian) Computing history and history of web in Iran highlights that the Farsi (Persian) language has had a significant role in moderating the diffusion of these technologies and how Iranians use such systems to inform their decision-making. We explored thus in a case from Iran, A.ir, an online shop whose market and revenue rely principally on its Google ranking and illustrate how a biographical view of the Iranian web-users' search habits in different time can improve quality of A.ir's data about its customers and its decision-making.

The A.ir study also illustrates how data creation, collection, modification and consumption may be viewed as shaping a network (Latour 2005) comprised of different actors competing to have their interests inscribed on the network. When these actors are from different societal groups, potential data quality issues can emerge as their societal context can evolve independently through time and such evolution should be considered during data consumption.

Reference

Akrich, M. 1992. "The De-Scription of Technical Objects," in *Shaping Technology/Building Society: Studies in Sociotechnical Change,* W.E. Bijker and J. Law (eds.). Cambridge, MA: MIT Press.

Ballou, D., Wang, R., Pazer, H., and Tayi, G.K. 1998. "Modeling Information Manufacturing Systems to Determine Information Product Quality," *Management Science* (44:4), pp. 462-484.

Cao, L., and Zhu, H. 2013. "Normal Accidents: Data Quality Problems in Erp-Enabled Manufacturing," *Journal of Data and Information Quality (JDIQ)* (4:3), p. 11.

Dadeh Pardazi Iran (dpi). "Our History." Retrieved 14 November, 2013, from http://www.dpi.ir/fa/companyintro/history

Dadkhah, K.M. 1985. "The Inflationary Process of the Iranian Economy, 1970-1980," *International Journal of Middle East Studies* (17:03), pp. 365-381.

Esfahbod, B. 2004. "Persian Computing with Unicode," *25th Internationalization and Unicode Conference*, Washington, DC.

Feyzi, K. 1985. "An Overview to the Adoption of Informatic Technologies from Beginning to the End of the 1970s in Iran," *Management Studies* (1:1), pp. 28-35.

Fleck, J. 1994. "Learning by Trying: The Implementation of Configurational Technology," *Research policy* (23:6), pp. 637-652.

Fleck, J. 1999. "Learning by Trying: The Implementation of Configural Technology," in *The Social Shaping of Technology,* D. MacKenzie and J. Wajcman (eds.). Berkshire: Open University Press.

Kopytoff, I. 1986. "The Cultural Biography of Things: Commoditization as Process," in *The Social Life of Things: Commodities in Cultural Perspective,* A. Appadurai (ed.). Cambridge: Cambridge University Press, pp. 64-91.

Latour, B. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

Lave, J. 1993. "The Practice of Learning," in *Understanding Practice : Perspectives on Activity and Context,* S. Chaiklin and J. Lave (eds.). Cambridge: Cambridge University Press.

Law, J. 1987. "Technology and Heterogeneous Engineering: The Case of Portuguese Expansion," in *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology,* W.E. Bijker, T.P. Hughes and T.J. Pinch (eds.). Cambridge, MA: MIT Press, pp. 1-134.

Law, J., and Singleton, V. 2005. "Object Lessons," *Organization* (12:3), pp. 331-355.

Lazer, D.M., Kennedy, R., King, G., and Vespignani, A. 2014. "The Parable of Google Flu: Traps in Big Data Analysis,").

Lee, Y., Chase, S., Fisher, J., Leinung, A., McDowell, D., Paradiso, M., Simons, J., and Yarsawich, C. 2007. "Ceip Maps: Context-Embedded Information Product Maps,").

Lee, Y.W. 2004. "Crafting Rules: Context-Reflective Data Quality Problem Solving," *Journal of Management Information Systems* (20:3), pp. 93-119.
Lee, Y.W., and Strong, D.M. 2003. "Knowing-Why About Data Processes and Data Quality," *Journal of Management Information Systems* (20:3), pp. 13-39.
Leonardi, P.M. 2010. "Digital Materiality? How Artifacts without Matter, Matter," *First Monday* (15:6).
Leonardi, P.M., and Barley, S.R. 2008. "Materiality and Change: Challenges to Building Better Theory About Technology and Organizing," *Information and Organization* (18:3), pp. 159-176.
Levitin, A.V., and Redman, T.C. 1993. "A Model of the Data (Life) Cycles with Application to Quality," *Information and Software Technology* (35:4), pp. 217-223.
Lloyd, A. 2005. "The Grid and Crm: From 'If' to 'When'?," *Telecommunications Policy* (29:2), pp. 153-172.
Marcus, G.E. 1995. "Ethnography in/of the World System: The Emergence of Multi-Sited Ethnography," *Annual review of anthropology* (24:1), pp. 95-117.
Monteiro, E., and Rolland, K.H. 2012. "Trans-Situated Use of Integrated Information Systems," *Eur J Inf Syst*), 03/20/online.
Orlikowski, W.J. 2000. "Using Technology and Constituting Structure: A Practice Lens for Studying Technology in Organizations," *Organization Science* (11:4), 2000, pp. 404-428.
Orlikowski, W.J. 2007. "Sociomaterial Practices: Exploring Technology at Work," *Organization studies* (28:9), pp. 1435-1448.
Parhami, B. 1997. "An Overview to the 40 Years History of Computer in Iran," *Gozaresh Computer* (138).
Pipino, L.L., Lee, Y.W., and Wang, R.Y. 2002. "Data Quality Assessment," *Communications of the ACM* (45:4), pp. 211-218.
Redman, T.C. 1998. "The Impact of Poor Data Quality on the Typical Enterprise," *Communications of the ACM* (41:2), pp. 79-82.
Sadoughi, R. 1981. "History of Informatics and Computer Systems in Iran (1)," in: *Ettelaat*. Tehran: Ettelaat Publishing Company.
Shenkar, O. 2001. "Cultural Distance Revisited: Towards a More Rigorous Conceptualization and Measurement of Cultural Differences," *Journal of international business studies*), pp. 519-535.
Sousa, C.M., and Bradley, F. 2005. "Global Markets: Does Psychic Distance Matter?," *Journal of Strategic Marketing* (13:1), pp. 43-59.
Sousa, C.M.P., and Bradley, F. 2006. "Cultural Distance and Psychic Distance: Two Peas in a Pod?," *Journal of International Marketing* (14:1), pp. 49-70.
Strong, D.M., Lee, Y.W., and Wang, R.Y. 1997. "Data Quality in Context," *Communications of the ACM* (40:5), pp. 103-110.
Torchiano, M., Di Penta, M., Ricca, F., De Lucia, A., and Lanubile, F. 2011. "Migration of Information Systems in the Italian Industry: A State of the Practice Survey," *Information and Software Technology* (53:1), pp. 71-86.
Walsham, G., and Sahay, S. 1999. "Gis for District-Level Administration in India: Problems and Opportunities," *MIS quarterly*), pp. 39-65.
Wand, Y., and Wang, R.Y. 1996. "Anchoring Data Quality Dimensions in Ontological Foundations," *Communications of the ACM* (39:11), pp. 86-95.
Wang, P. 2010. "Chasing the Hottest It: Effects of Information Technology Fashion on Organizations," *MIS quarterly* (34:1), pp. 63-85.
Wang, R.Y., Strong, D.M., and Guarascio, L.M. 1996. "Beyond Accuracy: What Data Quality Means to Data Consumers," *J. of Management Information Systems* (12:4), pp. 5-33.
Watts, S., Shankaranarayanan, G., and Even, A. 2009. "Data Quality Assessment in Context: A Cognitive Perspective," *Decision Support Systems* (48:1), pp. 202-211.
Williams, R., and Pollock, N. 2012. "Research Commentary—Moving Beyond the Single Site Implementation Study: How (and Why) We Should Study the Biography of Packaged Enterprise Solutions," *Information systems research* (23:1), pp. 1-22.
Williams, R., Stewart, J.K., and Slack, R.S. 2005. *Social Learning in Technological Innovation: Experimenting with Information and Communication Technologies*. Cheltenham: Edward Elgar Publishing.
Zhu, H., Madnick, S., Lee, Y., and Wang, R. 2012. "Data and Information Quality Research: Its Evolution and Future." Retrieved 30 July 2013, 2013, from http://web.mit.edu/smadnick/www/wp/2012-13.pdf